Running head: SOCIAL CONTEXT AND CYCLIC EFFECT ON WOMEN'S VOICES

Content matters: Cyclic effects on women's voices depend on social context

Wilhelm K. Klatt, Boris Mayer, & Janek S. Lobmaier

University of Bern

Author Note

Correspondence concerning this article should be addressed to Janek S. Lobmaier, Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland.

E-mail: janek.lobmaier@psy.unibe.ch

Abstract

Women's voices reportedly sound more attractive during the fertile days compared to the non-fertile days of their menstrual cycle. Here we investigated whether the speech content modulates the cyclic changes in women's voices. We asked 72 men and women to rate how interested they were in getting to know the speaker based on her voice. Forty-two naturally cycling women were recorded once during the late follicular phase (high fertility) and once during the luteal phase (low fertility) while speaking sentences of neutral and social content. Listeners were more interested in getting to know the speakers when hearing sentences with social content. Furthermore, raters were more interested in getting to know the speakers when these were recorded in the late follicular than in the luteal phase, but only in sentences with social content. Notably, levels of reproductive hormones (EP ratio) across the cycle phases did not significantly predict the preference for late follicular voices, but echoing the perceptual ratings, there was a significant EP ratio x speech content interaction. Phonetic analyses of mean fundamental frequency (F0) revealed a main effect of menstrual cycle phase and speech content but no interaction. Employing an action-oriented task, the present study extends findings of cycle-dependent voice changes by emphasising that speech content critically modulates fertility effects.

Keywords: Women's speech; mate preference; voice attractiveness; fertility; menstrual cycle, reproductive hormones

1. Introduction

The human voice is the most important and most expressive medium for interpersonal communication. The voice allows other people to form impressions of a speaker's personality (Eckert and Laver, 1994): It conveys rich dynamic information about the speaker's sex, emotional state, age, health, regional origin, competence, trustworthiness, and attractiveness (Schweinberger et al., 2014). Indeed, people can readily differentiate between voices they find attractive and voices they find less attractive. Interestingly, voice attractiveness correlates with other dimensions related to reproductive potential and health (Shoup-Knox and Pipitone, 2015), such as body symmetry (Hughes et al., 2002), facial attractiveness (Collins and Missing, 2003; Feinberg et al., 2005), body mass index (Wheatley et al., 2014), body attractiveness, and sexual behaviour (Hughes et al., 2004).

An evolutionary approach to attractiveness proposes that people should generally prefer individuals as mates who signal high reproductive health and fertility. The human voice is one of these cues to reproductive potential and fitness. For example, voices of women of reproductive age (19-30 years old) are rated as being more attractive and more feminine than voices of non-fertile age groups (11-15 and 50-65 years old; Röder et al., 2013). From this perspective, it stands to reason that the voice plays an important role in mate choice (Abend et al., 2015), since it can provide additional information about potential mates when visual cues are ambiguous or not available, for example in dimmed light or when a speaker is out of sight (Hughes et al., 2002; Pipitone and Gallup, 2008). Men reportedly have stronger preferences for higher pitched voices in women, reflecting femininity, whereas women seem to prefer lower pitched voices in men, reflecting masculinity (Collins and Missing, 2003; Jones et al., 2010). Consistent with this finding, women reportedly speak in a higher pitch to attractive men (Fraccaro et al., 2011) and men lower their voice pitch when speaking to attractive women (Hughes et al., 2010), suggesting that the voice might signal mating motivation. Jones et al. (2008) investigated vocal attractiveness depending on voice pitch and speech content. In

a two-alternative forced choice paradigm, they found that men preferred women's voices with raised pitch (manipulation equivalent to + 20 Hz). Notably, this preference was stronger when speech content signalled social interest in the listener ("I really like you") than when speech content signalled disinterest ("I don't really like you"). Female listeners also showed a preference for women's voices with raised pitch but this was not modulated by cues of social interest. Jones et al. (2008) thus concluded that voice preferences do not only depend on physical characteristics but also on speech content.

Accumulating evidence suggests that women's mating motivation varies across the menstrual cycle. For example, on days when fertility is high women report a greater desire to go to parties where they might meet men (Bullivant et al., 2004; Haselton and Gangestad, 2006), possibly because women's sexual desire is higher during the fertile phase (e.g., Roney and Simmons, 2013, 2016; Shirazi et al., 2018; van Stein et al. , 2019; for a review see Jones et al., 2019). Furthermore, women have been reported to dress (Haselton et al., 2007), walk and dance (Fink et al., 2012) more attractively in the fertile days of their cycle. An adaptationist interpretation of these cycle-dependent behavioural changes is that they facilitate and foster reproduction in such way that women present subtle cues of fertility to which observers respond (e.g., Haselton and Gildersleeve, 2016).

A number of studies suggest that also women's voices change during their cycle. Pipitone and Gallup (2008) recorded the voices of naturally cycling women and women taking hormonal contraceptives while they were counting from 1 – 10 at four different times across their cycle. In naturally cycling women, ratings of voice attractiveness on a 100-point unlabelled scale increased as the conception probability increased. In women taking hormonal contraception, there were no cycle effects on voice attractiveness ratings. Similarly, Banai (2017) phonetically compared the voices of naturally cycling women with women taking hormonal contraceptives while producing sustained vowels. Phonetic analyses showed that in women with natural menstrual cycles, minimum fundamental frequency (F0) was

significantly higher in the late follicular phase than during menstruation and the luteal phase. Moreover, voice intensity (perceived loudness) was lower in the luteal phase than during menstruation and the late follicular phase. In women using hormonal contraception, there were no voice changes across the cycle. Bryant and Haselton (2009) recorded voices of naturally cycling women when their fertility was low and when their fertility was high. They found a significantly higher voice pitch in high-fertility recordings than in low-fertility recordings, but only in meaningful speech (i.e., when speaking the introductory sentence "Hi, I'm a student at UCLA"). In vowels, there was no effect of current fertility. This led the authors to conclude that cycle-dependent vocal changes may occur during social communication only. However, it should be noted that some studies did not observe effects of menstrual cycle on phonetic characteristics. Barnes and Latman (2011) analysed the voices of naturally cycling and women on hormonal birth control phonetically at five time points throughout the cycle while speaking a custom-designed sentence. In eight acoustic parameters, they found no significant differences across the menstrual cycle. In addition, there were no significant differences between women with natural cycles and women taking hormonal contraceptives. Likewise, Meurer et al. (2009) failed to find phonetic differences between follicular and luteal phase recordings while uttering a vowel, speaking a meaningless sentence, or a neutral sentence ("I will go to Gramado during my winter holidays"). Fischer et al. (2011) combined perceptual ratings with phonetic analyses based on daily voice recordings throughout the menstrual cycle. In free speech, F0 and variation in F0 increased prior to ovulation and showed a distinct drop on the day of ovulation. In line with these phonetic analyses, men rated vocal attractiveness to be higher during the pre-ovulatory period than on the day of ovulation itself (in a two-alternative forced choice). Employing a speed dating paradigm, Karthikeyan and Locke (2015) recorded naturally cycling women while responding to attractive male voices. Men perceived women's responses to be more attractive on days when women's fertility was high (on a 100-point scale, labelled "least attractive"/"most

attractive"). Against the authors' expectation, however, women spoke in a lower voice pitch during the fertile window than during non-fertile phases. Shoup-Knox and Pipitone (2015) found that voices of naturally cycling women were not only rated as more attractive at high fertility but also elicited a more pronounced increase in galvanic skin response compared to low fertility voices, both in men and women. Increased galvanic skin responses are typically interpreted as a sign of heightened arousal. Together, the findings of Shoup-Knox and Pipitone (2015) suggest that the alleged cues to fertility in voices not only affect conscious responses but also provoke reactions of the autonomous nervous system.

One possible explanation for cyclic shifts in voice attractiveness may be found in the fact that levels of women's reproductive hormones (e.g., estrogen and progesterone) vary over the menstrual cycle. Changes in hormonal secretion cause histological changes in the mucus and glandular cells in the larynx, which may result in voice changes (Amir et al., 2002). Further, specific receptors for sex hormones within the vocal fold mucosa have been found (Schneider et al., 2007). Puts et al. (2013) provided evidence for the relationship between levels of reproductive hormones and voice changes. They found that cycle-dependent shifts in progesterone and estradiol predicted how attractive women's voices were rated when reading a passage from an articulation drillbook. This explanation – which implies a link between hormone fluctuations and voice changes – suggests that cycle-dependent voice changes should be observable in all voice samples, irrespective of speech content.

An alternative explanation for the finding that women's voices sound more attractive when recorded during the fertile period of their cycle is that women experience increased mating motivation. As a result, women might speak with a more attractive voice on days when they are fertile. This might especially be the case when primed with a mating context (e.g., Karthikeyan and Locke, 2015), implying that fertility-dependent voice changes need a social trigger to unfold.

The above-mentioned studies used different methods to track the menstrual cycle of the speakers. Pipitone and Gallup (2008) and Puts et al. (2013) relied on counting methods, Fischer et al. (2011) used hormone assays measuring metabolites of estrogen and progesterone in urine samples, and Karthikeyan and Locke (2015) used digital ovulation kits measuring the metabolite of luteinizing hormone in urine. These techniques differ in accuracy and it is conceivable that at least some speakers in these studies were tested in a wrong cycle phase or during an anovulatory cycle (for a discussion, see Lobmaier and Bachofner, 2018). In addition, the voices in the above mentioned studies were recorded within different contexts. Some researchers employed a dating scenario where women left a voice message for an attractive man (Fraccaro et al., 2011; Karthikeyan and Locke, 2015), others recorded a voice message for a phone survey (Hughes et al., 2010), a standard voice passage taken from an articulation drill book (Puts et al., 2013), a neutral phrase (Wells et al., 2013), free speech (e.g., comments on the weather or plans for the day; Fischer et al., 2011), counting from 1 to 10 (Pipitone and Gallup, 2008), or sustained vowels (Banai, 2017; Collins and Missing, 2003; Feinberg et al., 2005; Jones et al., 2010). Obviously, some of these recording scenarios have a more "social" framing than others, which could have affected women's motivation to sound more attractive. The present study set out to directly test whether the degree of social framing modulates menstrual cycle effects on women's voices. It is conceivable that especially during social communication, menstrual cycle phase would have an effect on a woman's voice.

In the present study, the voice of naturally cycling women was recorded during their late follicular and during their luteal phase. We used two different levels of speech content: Speakers were recorded while uttering sentences of neutral content (e.g., "The ship is arriving at the harbour") and "social" sentences framed in a dating context (e.g., "Hello, may I invite you to a cup of coffee?"). We thus manipulated the social framing of the sentences, enabling us to directly compare women's voices during social speech (implying the possibility of meeting and getting to know a person) and during non-social speech. If women show

increased mating motivation in the fertile window, we would expect vocal shifts to occur primarily in social sentences (i.e., in a getting-to-know context), but not in neutral sentences. Alternatively, if vocal changes across the menstrual cycle can be attributed to direct hormonal influences on the vocal apparatus (cf. Amir et al., 2002; Schneider et al., 2007), vocal shifts should occur irrespective of speech content, that is, both in neutral and social sentences.

We asked how much listeners would want to get to know the speaker based on her voice. By doing so, the judgment was action-oriented and personally relevant to the listeners, as this question implies that they would actually get to meet the speakers. Since women and men may have different motivations to meet the speakers, we included men and women as raters. Additionally, to understand the physical properties of potential voice changes, recordings were analysed phonetically using Praat software (Boersma and Weenink, 2018). In contrast to studies in which conception probability was only calculated using forward-counting methods (Pipitone and Gallup, 2008; Puts et al., 2013), we determined and verified the menstrual cycle phases by hormone tests based on urine and saliva.

Based on previous studies on perceived voice attractiveness, we predicted that listeners would show a higher interest in getting to know the speakers when their voices had been recorded during the fertile window compared to during the non-fertile luteal phase. With respect to speech content, we expected menstrual cycle effects to be more pronounced in sentences with social content than in neutral sentences, because fertile women may be particularly motivated to sound attractive when speaking sentences with social content (i.e., sentences implying the possibility of a personal meeting). Finally, we expected that men would show a larger preference than women to meet a currently fertile speaker, because from an evolutionary perspective it would seem particularly relevant for men to detect subtle cues to fertility.

2. **Material and methods**

2.1 Participants

Eighty-three speakers (all women) and 72 raters (36 women, 36 men) were recruited for this study via online advertisements, flyers, and the participant pool of the Institute of Psychology at the University of Bern. All participants provided written informed consent to participate. This study was approved by the local ethics committee and participants were treated in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Speakers were compensated either with course credits or 50 US$, raters were compensated with course credits or 20 US$.

2.2 Procedure

2.2.1 Voice recordings

Initially, all interested women were asked to complete an online survey with questions regarding age, smoking habits, mother tongue, reading disabilities, hearing problems, sexual orientation, relationship status, hormonal contraception, pregnancy, onset of last menstruation, regularity and length of menstrual cycle. To take part in this study, speakers had to meet following inclusion criteria: No use of hormonal contraceptives ("pill", "morning-after pill", hormonal contraceptive coil, contraceptive implants), no pregnancy, and no breastfeeding within the last three months; regular menstrual cycle (23–35 days in length), proficient in German (mother tongue), heterosexual orientation, no dyslexia, no hearing problems, no chronic smoking (more than 20 cigarettes a week). Eighty-three women who met the inclusion criteria were contacted by phone by a female research assistant who gave them more detailed information about the procedure.

To determine time of highest fertility, participants completed a series of urine tests measuring a metabolite of luteinizing hormone (LH) using one-step urine LH tests with a reported sensitivity of 10 mIU/ml (David One Step Ovulation Tests, Runbio Biotech, China, http://www.runbio-bio.com). Women were instructed to perform urine tests twice a day

(morning and evening) starting three days before the date of predicted peak fertility (based on the average cycle length of each individual woman using forward and backward-counting method). After a positive test result, participants continued performing LH tests until the results became negative for two subsequent days. Participants photographed each test using their smartphones and sent the picture to the research assistant, who verified whether the test was positive or not.

The women were either scheduled to be recorded approximately two days before the calculated day of peak fertility and again seven days after a positive LH test result (late follicular–luteal group) or they were scheduled seven days after the LH surge (luteal–late follicular group). Speakers of the luteal–late follicular group performed LH tests again in the following cycle and where scheduled to be recorded two days before the calculated day of peak fertility. Thus, LH tests were used to determine peak fertility and to verify that the cycle was ovulatory. Late follicular recording sessions took place between 4 days before and 24 hours after the LH surge, luteal phase recording sessions took place 6 to 12 days after the LH surge (see Table 1 for an overview of the time of recording relative to the LH surge). Order of recording sessions was counterbalanced across speakers.

--- Table 1 about here ---

At each test session, speakers provided saliva samples from which estradiol and progesterone levels were determined to confirm that the women were recorded in the right cycle phase (late follicular, luteal). We also assessed testosterone and cortisol levels, but these were not used in any analyses. Participants were instructed to refrain from eating and drinking anything but water for at least 30 minutes prior to saliva collection. Samples were collected by passive drool using a commercially available sampling device (SaliCaps, IBL International, Hamburg, Germany). The saliva samples were stored at –28°C and were later

analysed by an independent laboratory (Dresden Lab Service GmbH, Dresden, Germany) using liquid chromatography with coupled tandem mass spectrometry (LC-MS/MS). LC-MS/MS has become the method of choice for steroid analysis because of its high sensitivity, better reproducibility, greater specificity, and ability to analyse multiple steroids simultaneously. Both recording sessions took place at the same time of day between 8 and 11 AM.

The recordings were carried out in a soundproof booth under standardized conditions. A beyerdynamic MC 930 condenser microphone with a popkiller and 48 V phantom power was placed about 20 centimetres away from the speaker's mouth and connected to a Zoom H4n digital audio recorder (uncompressed WAV, 48 kHz, 16-bit sampling rate). Before starting the experiment proper, speakers were asked to read out a short newspaper article about voice research presented on a computer screen to warm up their voices and to adjust the recording level to −12 to −6 dB. The instruction to read out presented written sentences in a natural manner was given verbally and onscreen. Each recording session consisted of two blocks. In the first block, three sentences of affectively neutral content ("The ship is arriving at the harbour", "The key opens the door", "The doctor calls for the nurse") were presented. In the second block, three sentences with social content ("Hello, may I give you my phone number?", "Hello, may I invite you to a cup of coffee?", "Hello, I would like to meet you") were presented. Sentences were displayed on a laptop screen (Arial, 25 pt) using PsychoPy software (Peirce, 2007). The screen was placed in front of the speaker at a distance of approximately 50 centimetres. Sentences were presented separately in randomized order for five seconds each. After three practice trials with sentences that were not used subsequently, the experimenter left the booth and the speaker started the first block of the recordings (neutral sentences). Prior to the second block (social sentences), speakers were instructed to picture themselves in a situation in which they want to meet a person. Both test sessions

followed the exact same procedure except that after the second session participants were fully debriefed. Each recording session took about 20 minutes to complete.

Of the 83 women who initially entered the study, some had to be excluded because their recordings were unusable due to misspeaking (i.e., they spoke non-existent or other words than displayed, based on subjective judgments of an assistant blind to the hypotheses of the study; $N = 14$), because they had anovulatory cycles during the recording period (i.e., no LH surge, $N = 8$), did not conduct the required LH tests during the peri-ovulatory period ($N = 3$), were tested in the wrong cycle phase (too early/too late as revealed by LH tests, $N = 8$), or dropped out due to personal reasons ($N = 8$). Excluded speakers were neither rated perceptually nor analysed phonetically. Thus, the final sample consisted of 42 women between 18 and 35 years of age ($M = 22.9$, $SD = 3.8$). Twenty-one women completed the first session at high fertility and their second session in the luteal phase (late follicular–luteal group), 21 women were recorded first during the luteal phase and then during high fertility (luteal–late follicular group). All of them reported to be healthy (no mental and/or physical diseases) and not to have a hoarse voice, cough, or nasal congestion on the days of recording.

A total of 504 uncompressed WAV voice samples (42 speakers × 2 menstrual cycle phases × 6 sentences) were cut from raw recordings using Audacity® software (AudacityTeam, 2015). These voice samples were RMS (root mean square) normalised using Praat software (Boersma and Weenink, 2018) before they were presented to the raters (see next section).

2.2.2 Voice rating

Raters consisted of 36 women ranging in age between 19 and 40 years ($M = 23.6$, $SD = 4.1$) and 36 men ranging in age between 17 and 32 years ($M = 23.8$, $SD = 3.3$). Raters were tested individually in a quiet room. All reported to be proficient in German (mother tongue), to be of heterosexual orientation, to have no mental and/or physical diseases and no hearing

problems. PsychoPy software (Peirce, 2007) was used to present voice recordings via Sennheiser HD 201 headphones and raters were asked: "Based on her voice, how much would you like to get to know the speaker?". Voice samples were presented in fully randomized order, that is, neutral and social sentences were presented intermixed. Responses were given on a 100-point visual analogous scale with 1 being "not at all" and 100 being "very much so". Each recording was presented only once and the next trial started when the participants confirmed their rating. After the experiment, which took approximately 45 minutes to complete, raters were debriefed, compensated, and thanked for participation. None of the participants indicated having been aware of the purpose of the study.

2.3 Statistical analyses

Statistical analyses were performed using SPSS 25.0 and the R packages *lme4* (Bates, Maechler, Bolker, & Walker, 2015), *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017), *optimx* (Nash & Varadhan, 2011), and *pbkrtest* (Halekoh & Højsgaard, 2014). The level of significance was set at $p = .05$ for all analyses.

In a first step we calculated the intraclass correlation coefficient (ICC) based on a mean rating, absolute agreement, 2-way random effects model to quantify how much the raters agreed on their judgments of women's voices.

The main analysis includes two sets of models: In the first set, the principal dependent variable – voice ratings – is predicted by menstrual cycle phase and speech content, testing our main research question. Within this first set we ran three additional models (with EP ratio, logarithmized progesterone, and logarithmized estradiol as predictors of voice ratings in addition to speech content) in attempt to further elucidate the proposed effect of cycle phase by predicting voice ratings with hormonal variations across cycle phases. Because of the high intercorrelations of these predictors with cycle phase and with each other, they cannot be analyzed in a single model. To avoid Type I error inflation due to multiple testing in this set

of four models we used a Bonferroni-adjustment. Since we applied a semiparametric bootstrap for our final results (see below), the adjustment was made by using 98.75% confidence intervals instead of 95% confidence intervals for all effects in these four models, thereby correcting alpha to 0.05/4 = 0.0125. The second set of models refers to the prediction of five different phonetic parameters by menstrual cycle phase and speech content. Thus, five different (but partly related) dependent variables are predicted by the same set of independent variables. To avoid Type I error inflation due to multiple testing in this set of five models we use a Bonferroni-adjustment by reporting 99% confidence intervals instead of 95% confidence intervals for all effects in these five models (correcting alpha to 0.05/5 = 0.01).

We ran multilevel linear mixed regression models (Snijders & Bosker, 2012) with the rated interest in getting to know the speaker as the dependent variable. In all models, the Level-1 predictors were nested in speakers as well as raters as Level-2 subject variables. Since speakers and raters were not nested (as all raters rated all speakers), crossed random effects were defined (Baayen, Davidson, & Bates, 2008). We used an unstructured correlation matrix for the Level-2 residuals and aimed to include random slopes for all Level-1 predictors for both subject variables (in addition to random intercepts). To avoid over-parameterization and related estimation/convergence problems we determined the appropriate random effects structure for a specific model by starting with a model including all possible random slopes and tested their respective statistical significance via likelihood-ratio (LR) model comparisons using the *ranova* function from the *lmerTest* package. Non-significant random slope variance parameters (together with their associated covariance parameter(s)) were then removed from the model and the procedure was repeated with the reduced model. For example, when the model included two Level-1-predictors as well as their interaction, random slopes (separately for both subject variables rater and speaker) were first specified for the two predictors as well as their interaction. For such a model, the LR test for random (co)variance parameters separately tests the two random slope variances (for rater and speaker) for the highest order

term (i.e. the interaction of the two Level-1 predictors). In the case that one or both of these random slope variances and associated covariances appeared to be irrelevant for the model according to the LR test, a reduced model without one or both interaction random slopes was fitted and this model was again compared to models without a specific random slope (now of the Level-1 main effects) in separate LR tests. The procedure was stopped when all pertinent LR tests for a specific model were significant, indicating that the remaining random slope parameters needed to be included in the model (for details of the resulting random effects structures for the different analyses see below). Since in the current data set the main dependent variable regarding the rated interest in getting to know the speaker was clearly non-normally distributed for the subject variable rater, and less clearly so for the subject variable speaker we used bootstrap analyses for the fixed effects (see below). For consistency, we also ran all model comparisons for random effects components using the bootstrap procedure provided by the *pbkrtest* package (Halekoh & Højsgaard, 2014), applying 1000 bootstrap samples. To avoid unnecessary (because unequivocal) computing-intensive bootstrap model comparisons we only bootstrapped those model comparisons where the LR test's p value was equal to .001 or greater (see Table 2). Compared to the LR tests, the bootstrap analyses showed that for our primary analysis a random slope of menstrual cycle phase had to be included. For all other comparisons, LR tests and bootstrap analyses led to the same conclusion of significance/non-significance with regard to the respective random slope variance component.

This procedure worked well, except for the models with logarithmized progesterone and logarithmized estradiol as predictors of voice ratings. In these models, the respective two random slopes of the interaction term (of progesterone or estradiol, with the second predictor being speech content) were significant in LR tests, but the models including these random slopes did not reach convergence in the restricted maximum likelihood estimation (REML). Therefore, the random slopes of these interaction terms were removed from the model.

After determining the appropriate (or as in the last case: realizable) random effects structure all mixed models were fit with REML, using the *nlminb* optimizer from the *optimx* package (Nash & Varadhan, 2011) for enhanced convergence. Additionally, we used the semiparametric bootstrap procedure provided by the *bootmer* function of the *lme4* package using 1000 bootstrap samples and percentile bootstrap confidence intervals. In this bootstrap procedure the level-1 errors are sampled from the distribution of response residuals from the fitted model while the level-2 errors stay at their estimated values (bootstrapping the empirical level-2 residuals is not considered good practice since it results in systematically underestimated random effects variances; see Morris, 2002). Bootstrapping procedures are indicated when the normal theory assumption of parametric linear models is likely to be violated. In the case of mixed effects models, the regression residuals should follow a (multivariate) normal distribution at both levels of analysis. As these assumptions are often violated, bootstrapping is a robust alternative that can provide more accurate inferences (Fox, 2016). For many effects, the parametric and the bootstrap results coincided with regard to the significance/non-significance of effects, but there were also a number of effects where the REML and bootstrap results differed. We report the more robust bootstrap results in the text and present the parametric results (side by side with the bootstrap results for comparison) in Tables 2 and 4.

The results of the REML estimated mixed models and those of the semiparametric bootstrap differed particularly for our main analysis predicting the rated interest in getting to know the speaker with menstrual cycle phase and speech content, especially when correcting for multiple testing. As an additional validity check we therefore carried out a 2 (menstrual cycle phase) × 2 (speech content) repeated measures ANOVA with ratings aggregated within each rater × speaker combination and with both raters and speakers as person variables. The results of this analysis coincided with the mixed model bootstrap results for non-aggregated ratings (see Results section).

We report the intra-class coefficients from the respective intercept-only models to document the proportion of variability in the dependent variable that was due to the subject variable(s). All corresponding LR tests for the significance of the respective level-2 (subject) random effect variance component were highly significant ($p < 0.001$), indicating the necessity to employ hierarchical linear models. The only exception to this was the variance component of the subject variable speaker for the dependent variable maximum F0, where the intra-class correlation was relatively small and the random variance component was statistically detectable at the $p < 0.05$ level in the corresponding LR test. Tables 2 and 4 contain all parameter estimates, standard errors and confidence intervals for the fixed effects as well as the estimated random effects variance components together with their respective p values from LR tests and (if applicable) bootstrap model comparisons.

In the case of significant interactions (fixed effects) of the respective focal predictor with the proposed moderator variable speech content (0 = neutral; 1 = social), we report the simple slope of the focal predictor for *neutral* sentences (speech content = 0/reference category) from the model including the interaction term. For the simple slope of the respective focal predictor for *social* sentences we recoded the speech content variable (resulting in 0 = social/reference category) and fitted the model again. This allowed direct tests of conditional effects together with interaction effects in the same model, making use of the full sample rather than conducting separate analyses with subsamples for these effects.

The first model concerned the prediction of voice ratings by menstrual cycle phase and speech content as well as their interaction. We first computed a main effects model that included two indicator variables reflecting menstrual cycle phase (0 = luteal phase; 1 = late follicular phase) and speech content (0 = neutral; 1 = social) as level-1 predictors of voice ratings (fixed effects). We repeated the analysis after adding the menstrual cycle phase × speech content interaction as additional level-1 predictor. To test for differences of the obtained effects according to raters' gender, we additionally computed a model including a 3-

way interaction of the level-2-predictor gender and the two level-1 predictors. The random effects structure for the main effects model included random slopes for both menstrual cycle phase and speech content for both subject variables (in addition to random intercepts that were included in all analyses). Random slopes were statistically detectable at $p < .001$ in the LR tests with the exception of the menstrual cycle phase (MCP) random slope that became significant only in the bootstrap model comparison ($p = .038$, see Table 2). For the model including the interaction effect a random slope for the interaction term for speaker was added to the above random effects structure of the main effects model ($p < .001$) while for the subject variable rater no interaction random slope was necessary (bootstrap $p = .268$).

To test whether speakers' hormone levels might explain raters' interest in getting to know the speaker, we calculated the estradiol-to-progesterone (EP) ratio for each speaker separately for each cycle phase. The EP ratio is a reliable index for current fertility (Baird et al., 1991; Rehman et al., 2014; Roney, 2018). The first model included speakers' EP ratio and speech content (0 = neutral; 1 = social) as predictors of voice ratings (main effects model). As before, speakers and raters were entered as subject variables to define crossed random effects. We repeated the analysis after adding the EP ratio × speech content interaction as additional predictor. The random effects structure for the main effects model included random slopes of menstrual cycle phase and speech content for both subject variables (rater and speaker, all *p*s < .001, except for EP ratio | rater bootstrap $p = .009$, see Table 2). For the model including the interaction effect an additional random slope for the interaction term was needed for speaker ($p < .001$) but not for rater (bootstrap $p = .220$) according to the procedure described above.

Logarithmized progesterone and estradiol levels were used as additional predictors for the effect of hormone levels on voice ratings in separate models. As for the EP ratio, both variables had values for each cycle phase for each speaker. In one model, logarithmized progesterone was entered as Level-1 predictor for voice ratings together with speech content, and in the other model logarithmized estradiol was entered together with speech content. As

with all other analyses, the respective interaction with speech content was entered in an additional analysis. The realizable random effects structure for models with progesterone and with estradiol as predictor has been laid out above: both in the main effects models as well as in the interaction effects models random slopes of speech content as well as logarithmized progesterone or estradiol, respectively, were defined for the subject variable speaker (all *p*s < .001, except bootstrap $p = .011$ and $p = .010$ for logP | rater in the main effects/ interaction effect model, respectively, see Table 2).

Praat software (Boersma and Weenink, 2018) was used to analyse the voice samples for the following phonetic parameters: Mean F0 (corresponds to perceived voice pitch), variation in F0 (intonation, F0 SD), minimum F0, maximum F0 (pitch range), and variation in perceived loudness (Intensity SD). Phonetic analyses were performed automated based on Praat scripts, the frequency range was set to 100–600 Hz following Feinberg et al. (2006). Apart from this, default settings were used. We ran separate multilevel linear regressions with mean F0, F0 SD, minimum and maximum F0, and intensity SD as dependent variables. We first ran main effects models including the two indicator variables menstrual cycle phase (0 = luteal phase; 1 = late follicular phase) and speech content (0 = neutral; 1 = social) as Level-1 predictors of phonetic parameters (fixed effects). We repeated the analysis after adding the menstrual cycle phase × speech content interaction as additional Level-1 predictor. For all phonetic parameter models only speakers were defined as Level-2 subject variables. According to our procedure to determine the appropriate random effects structure, random slopes for menstrual cycle phase *and* speech content were estimated in the models for Mean F0 (MCP | speaker < .001 for both models; SC | speaker p = .001 for the main effects model and p = .002 for the interaction effects model, see Table 4). For the F0 SD models only a speech content random slope was necessary (bootstrap *p*s = .004/.002 for the main and interaction effects models, respectively) while the random slopes for menstrual cycle phase were non-significant (bootstrap *p*s = .521/.507 for the main and interaction effects models,

respectively). No random slope (only random intercept) was needed for F0 min models (for the bootstrap *p*s refer to Table 4). For the F0 max models again only random slopes for speech content were included (both *p*s < .001) while the random slopes for menstrual cycle phase were non-significant (bootstrap *p*s = .424/.448 for the main and interaction effects models, respectively, see Table 4). For the Intensity SD models random slopes for both predictors were included in both the main effects and the interaction effect models (all *p*s < .001, see Table 4). Finally, for the interaction effect models of all five dependent variables measuring phonetic parameters no random slopes for the interaction term of menstrual cycle phase × speech content was necessary (with bootstrap *p*s of 0.630, 0.055, 0.667, 0.188 and 0.130 for Mean F0, F0 SD, F0 min, F0 max, and Intensity SD, respectively, see Table 4) .

3. **Results**

3.1 Perceptual ratings of voice recordings

Inter-rater reliability: Intraclass correlation was high, indicating excellent reliability (*ICC* = .977; 95% CI [.964, .986]). This suggests that raters agreed highly on which speakers they were more and which ones they were less interested to meet.

--- Figure 1 about here ---

Figure 1 shows listeners' means ratings depending on speakers' menstrual cycle phase and speech content.

The intercept-only model with the ratings as dependent variable showed an Intercept of $\hat{\gamma}_{00} = 51.65$ ($SE = 2.09$) and random effect variance components of the intercept of $\hat{\sigma}^2_{\upsilon_{0R}} = 105.6$ (LR $p < .001$) for the subject variable raters and of $\hat{\sigma}^2_{\upsilon_{0S}} = 121.5$ (LR $p < .001$) for the subject variable speakers, together with a level-1 residual variance of $\hat{\sigma}^2_{\varepsilon} = 349.0$ (see Table 2). This results in an ICC_R = 0.18 for the subject variable raters and an ICC_S = 0.21 for

the subject variable speakers, indicating that around 20% of the variance of the rated interests in getting to know the speaker was between raters and speakers, respectively (and thus around 80 % of the variance occurred within raters and speakers).

--- Table 2 about here ---

The model including speaker's menstrual cycle phase and speech content as predictors revealed that cycle phase and speech content both predicted the listener's interest in getting to know the speaker. Listeners were more interested in getting to know the speaker when she was recorded during the late-follicular compared to the luteal phase, b = 0.79, bootstrap SE = 0.19, bootstrap 95% CI [0.44, 1.18], bootstrap 98.75% CI [0.30, 1.27], and were more interested in meeting the speaker when she spoke social as compared to neutral sentences with b = 2.36, bootstrap SE = 0.18, bootstrap 95% CI [2.00, 2.74], bootstrap 98.75% CI [1.88, 2.85]. The menstrual cycle phase × speech content interaction was also statistically detectable with b = 1.48, bootstrap SE = 0.36, bootstrap 95% CI [0.78, 2.20], bootstrap 98.75% CI [0.64, 2.38]. We followed up this interaction with conditional effects (simple slope) analyses focusing on speech content as a moderator of menstrual cycle phase. The conditional effect of menstrual cycle phase was significant for social sentences with b = 1.54, bootstrap SE = 0.26, bootstrap 95% CI [1.01, 2.05], bootstrap 98.75% CI [0.87, 2.23], but not for neutral sentences with b = 0.05, bootstrap SE = 0.26, bootstrap 95% CI [-0.47, 0.53], bootstrap 98.75% CI [-0.65, 0.65]. Finally, we tested whether gender moderated the reported effects. Since the 3-way interaction of menstrual cycle phase × speech content × gender was non-significant with b = 0.46, REML SE = 0.75, bootstrap SE = 0.73, REML CIs: 95% [-1.02, 1.94], 98.75% [-1.43, 2.34], bootstrap CIs: 95% [-1.03, 2.01], 98.75% [-1.39, 2.50] (not in Table 2), we concluded that both female and male raters wanted to get to know the speaker slightly more when the speaker was in the late-follicular phase of her menstrual cycle compared to when she was in

the luteal phase, but only for the social content sentences recorded. As stated above, the conclusions with regard to the statistical detectability of effects were not the same when considering the REML CIs (see Table 2). Especially, the crucial MCP × SC interaction effect, but also the main effects were non-significant according to these CIs. Since we understand that such inconsistencies may raise questions with regard to the validity of our reported bootstrap results for this central analysis, we carried out additional repeated measures ANOVAs with both menstrual cycle phase and speech content as repeated factors, with aggregated ratings for each rater × speaker combination and with both raters and speakers as person variables. We report both the original RM-ANOVA results as well as additional semiparametric bootstrap CIs. RM-ANOVAS would be the "traditional way" of analyzing this kind of data where not all single data points are modelled as in our mixed-effects model analysis. The results of the repeated measures ANOVAs showed that the parameter estimates were essentially the same as those for the non-aggregated REML mixed-effects models, but with smaller standard errors. In particular, the menstrual cycle phase main effect was significant with $F(1, 11977) = 10.16$, $p = .001$, $p_{adj} = .006$ (bootstrap 95 % CI [0.31, 1.25], bootstrap 98.75 % CI [0.13, 1.38]), the speech content effect was significant with $F(1, 11977) = 90.13$, $p < .001$, $p_{adj} < .001$ (bootstrap 95 % CI [1.87, 2.84], bootstrap 98.75 % CI [1.74, 2.92]), the cycle phase × speech content interaction was significant with $F(1, 11977) = 8.87$, $p = .003$, $p_{adj} = .012$ (bootstrap 95 % CI [0.45, 2.56], bootstrap 98.75 % CI [0.15, 2.79]), and the cycle phase × speech content × gender interaction was non-significant with $F(1, 11977) = 0.213$, $p = .645$, $p_{adj} = 1$ (bootstrap 95 % CI [-1.53, 2.41], bootstrap 98.75 % CI [-2.36, 3.14]). Furthermore, the conditional effect of cycle phase for neutral sentences was non-significant with $p = .883$ and $p_{adj} = 1$ (bootstrap 95 % CI [-0.66, 0.76], bootstrap 98.75 % CI [-0.81, 0.90]), and that for social sentences was significant with $p < .001$ and $p_{adj} < .001$ (bootstrap 95 % CI [0.80, 2.29], bootstrap 98.75 % CI [0.58, 2.53]). Thus, all effects that were significant according to the bootstrap CIs of the non-aggregated mixed-effects models were also

statistically detectable in the repeated measures ANOVAs *F* tests as well as according to the bootstrap CIs of these models. We consider this an additional indication for the robustness of the bootstrap results and suspect that especially the non-normal distributions of the ratings within each rater × speaker combination together with the fact that there were only six observations (sentences) for each of these combinations contributed to inflated standard errors in the REML mixed-effects models including all pertinent random slopes.

Regarding effect size, the effect of 1.54 for social sentences corresponded to 0.10 standard deviations of the voice ratings which was *SD* = 15.18 (mean of the menstrual-cycle-phase-pooled SDs for all raters × speaker combinations), and can thus be considered a rather small effect. Figure 2 illustrates the variability of the conditional effects of menstrual cycle phase across speakers.

--- Figure 2 about here ---

3.2 Hormone assays

Hormone levels during the late follicular and luteal phase are shown in Table 3. A Kolmogorov-Smirnov test indicated that hormonal data were not normally distributed. Hence, nonparametric Wilcoxon signed-rank tests were used to compare the hormone levels between both cycle phases. These analyses revealed that progesterone levels were significantly higher in the luteal phase than in the late follicular phase ($Z = -4.585$, $p < .001$). Levels of estradiol ($Z = -.961$, $p = .34$), testosterone ($Z = -.487$, $p = .63$), and cortisol ($Z = -1.500$, $p = .13$) did not differ between the two phases. EP ratio was significantly higher in the late follicular phase than in the luteal phase ($Z = -3.858$, $p < .001$).

---Table 3 about here---

The mixed effects model bootstrap analysis revealed that the effect of speakers' EP ratio on voice ratings was non-significant with *b* = 2.06, bootstrap *SE* = 0.94, bootstrap 95% CI [-1.06, 2.66], bootstrap 98.75% CI [-1.64, 3.21]. The effect of speech content - the second predictor also in this model - significantly predicted voice ratings which were higher for sentences with social content as compared to neutral sentences with *b* = 2.36, bootstrap *SE* = 0.19 , bootstrap 95% CI [2.00, 2.74], bootstrap 98.75% CI [1.88, 2.86]. Notably, the EP ratio × speech content interaction was statistically detectable with *b* = 5.95, bootstrap *SE* = 1.15, bootstrap 95% CI [1.95, 6.28], bootstrap 98.75% CI [1.25, 7.22]. Conditional effects (simple slope) analyses focusing on speech content as a moderator of the EP ratio effect revealed that the conditional effect of EP ratio was non-significant for both social and neutral sentences with *b* = 5.53, bootstrap *SE* = 1.26 , bootstrap 95% CI [0.58, 5.59], bootstrap 98.75% CI [-0.10, 6.44], and with *b* = -0.41, bootstrap *SE* = 0.87, bootstrap 95% CI [-2.65, 0.73], bootstrap 98.75% CI [-3.20, 1.16], respectively (see Table 2).

Additionally, models were calculated with logarithmized progesterone as well as estradiol levels, respectively, as predictors of voice ratings. Results showed that the effect of speakers' progesterone on voice ratings was non-significant with *b* = -0.18, bootstrap *SE* = 0.29, bootstrap 95% CI [-0.36, 0.79], bootstrap 98.75% CI [-0.50, 0.93]. The effect of speech content in this model was significant with *b* = 2.36, bootstrap *SE* = 0.19, bootstrap 95% CI [2.00, 2.74], bootstrap 98.75% CI [1.88, 2.86]. The main effects were again qualified by a significant progesterone × speech content interaction with *b* = -1.79, bootstrap *SE* = 0.38, bootstrap 95% CI [-2.56, -1.03], bootstrap 98.75% CI [-2.76, -0.89]. Conditional effects analyses focusing on speech content as a moderator of the progesterone effect showed that the conditional effect of progesterone was non-significant for social sentences with *b* = -1.11, bootstrap *SE* = 0.35, bootstrap 95% CI [-1.44, -0.06], bootstrap 98.75% CI [-1.61, 0.20], but was significant for neutral sentences with *b* = 0.67, bootstrap *SE* = 0.34, bootstrap 95% CI [0.42, 1.73], bootstrap 98.75% CI [0.25, 1.94] (see Table 2).

In the models with logarithmized estradiol as predictor of voice ratings, the main effect of estradiol was not significant with *b* = -0.34, bootstrap *SE* = 1.21, bootstrap 95% CI [-2.73, 1.95], bootstrap 98.75% CI [-3.39, 2.73], while the content main effect was again significant with *b* = 2.36, bootstrap *SE* = 0.19, bootstrap 95% CI [2.01, 2.74], bootstrap 98.75% CI [1.88, 2.85]. The interaction effect estradiol × speech content was also significant with *b* = -4.38, bootstrap *SE* = 1.48, bootstrap 95% CI [-7.75, -1.95], bootstrap 98.75% CI [-8.44, -1.30]. Both conditional effects of estradiol on the voice ratings were not statistically detectable with *b* = -2.62, bootstrap *SE* = 1.44, bootstrap 95% CI [-5.77, -0.12], bootstrap 98.75% CI [-6.68, 0.72] for social sentences, and with *b* = 1.75, bootstrap *SE* = 1.39, bootstrap 95% CI [-0.76, 4.73], bootstrap 98.75% CI [-1.44, 5.63] for neutral sentences (see Table 2).

3.3 Phonetic analyses of voice recordings

Mean fundamental frequency (F0): The ICC for mean fundamental frequency (F0) was 0.57, indicating that 57% of variation in this variable occurred between speakers. The model including speaker's menstrual cycle phase and speech content as predictors revealed that cycle phase and speech content both significantly predicted fundamental frequency (mean F0). Specifically, women showed a higher mean F0 in late-follicular than in luteal voice samples with *b* = 2.93, bootstrap *SE* = 1.07, bootstrap 95% CI [0.84, 5.03], bootstrap 99% CI [0.01, 5.48], and a higher mean F0 in neutral than in social sentences with *b* = -3.67, bootstrap *SE* = 1.08, 95% CI [-5.91, -1.65], 99% CI [-6.60, -1.13]. The menstrual cycle phase × speech content interaction was not significant with *b* = 2.97, bootstrap *SE* = 2.14, bootstrap 95% CI [-1.73, 7.26], bootstrap 99% CI [-2.8, 8.19] (see Table 4). Regarding effect size, the mean fundamental frequency differed 2.93 Hz across menstrual cycle phases corresponding to 0.23 standard deviations of the mean F0 variable (menstrual-cycle-phase-pooled *SD* = 12.73), constituting a small effect. Furthermore, it was 3.67 Hz lower for social as compared to

neutral sentences, corresponding to 0.28 standard deviations (speech-content-pooled *SD* = 12.89), also a small effect.

--- Table 4 about here ---

Variation in fundamental frequency (F0 SD): The ICC for variation in fundamental frequency (F0 SD) was 0.10, indicating that 10% of variation in this variable occurred between speakers, while 90% occurred within speakers. The model including speaker's menstrual cycle phase and speech content as predictors revealed that speech content, but not menstrual cycle phase, significantly predicted the variation in fundamental frequency (F0 SD). Specifically, women showed a higher F0 SD in neutral than in social sentences with *b* = -11.43, bootstrap *SE* = 1.56, bootstrap 95% CI [-14.32, -8.35], bootstrap 99% CI [-15.71, -7.30] . F0 SD did not differ between late-follicular and luteal voice samples with *b* = 0.69, *SE* = 1.57, bootstrap 95% CI [-2.38, 4.05], bootstrap 99% CI [-3.35, 4.96]. The menstrual cycle phase × speech content interaction was not significant with *b* = -2.37, bootstrap *SE* = 3.13, 95% CI [-8.38, 3.60], bootstrap 99% CI [-10.30, 5.95] (see Table 4). Regarding effect size, the variation in fundamental frequency was 11.43 Hz lower for social as compared to neutral sentences, corresponding to 0.65 standard deviations of the F0 SD variable (speech-content-pooled *SD* = *SD* = 17.49), indicating a medium-sized effect.

Minimum fundamental frequency (F0 min): The ICC for minimum fundamental frequency (F0 min) was 0.24, indicating that 24% of variation in this variable occurred between speakers (76% within speakers). The model including speaker's menstrual cycle phase and speech content as predictors minimum fundamental frequency (F0 min) revealed that both predictors were non-significant (menstrual cycle phase: *b* = 0.68, bootstrap *SE* = 2.73, bootstrap 95% CI [-4.83, 5.78], bootstrap 99% CI [-6.59, 7.87]; speech content: *b* = -

6.32, bootstrap *SE* = 2.88, bootstrap 95% CI [-11.63, -0.91], bootstrap 99% CI [-13.11, 0.99]). The menstrual cycle phase × speech content interaction was also not significant with *b* = 5.41, bootstrap *SE* = 5.58, bootstrap 95% CI [-5.55, 16.30], bootstrap 99% CI [-9.83, 18.93] (see Table 4).

Maximum fundamental frequency (F0 max): The ICC for maximum fundamental frequency (F0 max) was 0.04, indicating that only 4% of variation in this variable occurred between speakers, while 96% occurred within speakers. The between-speaker variance component was nevertheless significant as already reported above [LR $\chi^2(1)$ = 4.21, *p* = 0.040]. The results for the model including speaker's menstrual cycle phase and speech content as predictors showed that speech content, but not menstrual cycle phase, significantly predicted the maximum fundamental frequency (F0 max). Specifically, women showed a higher F0 max in neutral than in social sentences with *b* = -48.93, bootstrap *SE* = 9.40, bootstrap 95% CI [-68.11, -30.06], bootstrap 99% CI [-72.85, -24.98]. F0 max did not differ between late-follicular and luteal voice samples, *b* = -3.36, bootstrap *SE* = 9.61, bootstrap 95% CI [-21.35, 17.47], bootstrap 99% CI [-26.86, 22.11]. The menstrual cycle phase × speech content interaction was not significant with *b* = -20.91, bootstrap *SE* = 19.15, bootstrap 95% CI [-57.57, 16.41], 99% CI [-67.60, 32.69] (see Table 4). With respect to effect size, the maximum fundamental frequency was 48.93 Hz lower for social as compared to neutral sentences, corresponding to 0.44 standard deviations (speech-content-pooled *SD* = 110.15) of the F0 max variable (small to medium-sized effect).

Variation in perceived loudness (intensity SD): The ICC for variation in perceived loudness (intensity SD) was 0.30, indicating that 30% of variation in this variable occurred between speakers (70% within speakers). The model including speaker's menstrual cycle phase and speech content as predictors revealed that speech content, but not menstrual cycle phase, significantly predicted the variation in intensity (intensity SD). Specifically, women showed a higher intensity SD in social than in neutral sentences with *b* = 0.52, bootstrap *SE* =

0.08, bootstrap 95% CI [0.36, 0.67], bootstrap 99% CI [0.31, 0.72]. Intensity SD did not differ between late-follicular and luteal voice samples with b = 0.04, bootstrap SE = 0.08, bootstrap 95% CI [-0.12, 0.18], bootstrap 99% CI [-0.15, 0.25]. The menstrual cycle phase × speech content interaction was also not statistically detectable, b = 0.03, bootstrap SE = 0.15, bootstrap 95% CI [-0.27, 0.32], bootstrap 99% CI [-0.34, 0.43] (see Table 4). Regarding effect size, the variation in perceived loudness was 0.52 dB higher for social as compared to neutral sentences, corresponding to 0.49 standard deviations (speech-content-pooled SD = 1.07) of the intensity SD variable (medium-sized effect).

For descriptive statistics of the phonetic analyses, see electronic supplementary material, Table S1.

4. **Discussion**

The present study investigated the role of social content in the cyclic variations of women's voices. Naturally cycling women were recorded when speaking neutral sentences and sentences with social content (i.e., implying a situation where the speaker wants to meet the listener), once during the late follicular phase and once during the luteal phase. Independent raters were asked how much they were interested in getting to know the speakers based on their voices. By manipulating the social framing of the sentences we could directly compare ratings of social speech with ratings of non-social speech. Three main findings emerged. Firstly, listeners of both sexes showed a higher interest in getting to know the speaker when her voice was recorded during the fertile days than during non-fertile days of her cycle. Secondly, listeners were more interested in getting to know the speakers when they spoke sentences with social content than when they spoke neutral sentences. Thirdly, and most interestingly, we found that listeners' higher willingness to get to know the fertile speakers was modulated by speech content: A significant interaction between speakers'

menstrual cycle phase and speech content revealed that only in social sentences, listeners had a higher interest to meet the fertile speakers.

Previous research suggests that voices of naturally cycling women are rated to be more attractive when the women were recorded during the fertile phase of their cycle (e.g., Pipitone and Gallup, 2008; Puts et al., 2013; Shoup-Knox and Pipitone, 2015). These findings are interpreted as such that, through their voice, women present subtle cues of fertility which are detectable by listeners. The present study extends these findings by using a more action-oriented measure; listeners were more willing to actually get to know the speaker when her voice was recorded during the late follicular phase than during the luteal phase. This suggests that the preference for fertile women's voices persists when using other assessment criteria than attractiveness evaluations.

In the present study, we manipulated the social content of the sentences that had to be rated. Three sentences were of neutral content (e.g., "The ship is arriving at the harbour") and three sentences clearly had a social content (e.g., "Hello, may I invite you for a cup of coffee?"). Unsurprisingly, we found that listeners were more willing to get to know the speakers when they spoke sentences with social content than neutral sentences.

Interestingly, listeners showed a preference for fertile speakers only when these uttered sentences with a social content, implying that cyclic changes in women's voices occur primarily during social communication (cf. Bryant and Haselton, 2009; but see Pipitone and Gallup, 2008). In addition, these context-specific cycle effects were qualified by phase-specific hormone changes in estradiol and progesterone. The estradiol-to-progesterone ratio (EP ratio) across cycle phases did not significantly predict listeners' preferences for late follicular voices, but there was an interaction with speech content. However, both the positive effect for sentences with social content and the negative effect for sentences with neutral content did not reach significance after correcting for multiple testing. For progesterone levels, there was again a significant interaction with speech content, resulting in a statistically

detectable positive effect on the ratings for neutral sentences while a negative effect for social sentences was non-significant after correcting for multiple testing. Very similar results were obtained for estradiol levels: again, there was a statistically detectable interaction effect with speech content, but none of the conditional effects (with a negative sign for social sentences and a positive sign for neutral sentences) reached significance. The fact that speech content modulated the effect of current fertility (EP ratio) and progesterone levels on perceptual ratings renders the hypothesis unlikely that cyclic shifts in women's voices arise from hormonal changes directly acting on the vocal apparatus (Amir et al., 2002; Puts et al., 2013). If this hypothesis were true, cyclic hormonal changes should have a general effect on voice production and should not be modulated by speech content. Indeed, we found no indication that the speakers' EP ratios had a general influence on listeners' voice ratings (but see Puts et al., 2013 for a different view). Instead, it seems more likely that in the late follicular phase, women were more motivated to meet other people in general or men in particular (cf. Bullivant et al., 2004; Haselton and Gangestad, 2006; Karthikeyan and Locke, 2015), and that this motivation was conveyed only in sentences that implied the possibility of an actual meeting with the speaker.

A possible limitation is, because listeners were rating how much they would like to get to know the women, and because the women were making statements relevant to how much they would like to meet others, it would have been possible that listeners simply focused on any available cues of how sincere the speakers' social statements were, but in doing so neglected other voice features. In a future study it might be informative to have neutral sentences and social sentences rated separately by an independent sample. This will allow to test whether the missing effect of cycle phase in neutral sentences can be explained by raters listening more carefully to social sentences.

Another limitation refers to the rating task. In a future study it will be interesting to have the voice recordings to be rated for attractiveness or perceived sexiness in addition to

desire to getting to know the women, since the latter may be affected more by how friendly than how attractive the voice sounds.

We interpret the finding that men and women were more willing to meet currently fertile women, particularly when they uttered sentences that implied the possibility of a real meeting, as cyclic differences on the speakers' side (e.g., the women might have been more motivated to sound likeable or approachable). An alternative interpretation is that this effect might reflect higher motivation of the raters to listen carefully when judging sentences that implied meeting the speaker. Future work will need to disentangle whether the source of these cycle effects lies in the speaker or the listener, or both.

Our prediction that men would be more interested than women in getting to know fertile women was not met. Both male and female listeners showed more interest in getting to know the fertile speakers. This is in line with studies which found no difference between men and women when rating women's voices for attractiveness (Pipitone and Gallup, 2008; Shoup-Knox and Pipitone, 2015) and suggests that fertile women convey something in their voices that is equally interesting for men and women. Even though men and women may have different motivations to meet other women (pleasant encounter, friend, mate), it seems from the data in the present study that the specific motivations of the listeners did not modulate how much a listener wanted to get to know a speaker.

To investigate whether the cycle phase affects phonetic parameters of women's voices, we conducted phonetic analyses on the voice samples. These analyses revealed an effect of menstrual cycle on mean F0, suggesting that differences in perceptual ratings are reflected in voice pitch, the most prominent phonetic measure in human voices. In line with earlier studies (Bryant and Haselton, 2009; Fischer et al., 2011), we found that women spoke with a higher mean F0 during the high-fertility period compared to the low-fertility period. This is in contrast to Karthikeyan and Locke (2015) who found a lower voice pitch when fertile. In both studies, however, women's voices were rated to sound more attractive when fertile,

suggesting that mean F0 is not the main component of perceived vocal attractiveness. In the present study no cycle effect on minimum F0 or F0 variation was observed (Banai, 2017; Fischer et al., 2011). Our findings contrast research reporting no effect of menstrual cycle on phonetic characteristics of women's voices (Barnes and Latman, 2011; Meurer et al., 2009). Sentence content significantly predicted differences in all phonetic parameters (with the exception that the effect on minimum F0 was not statistically detectable after correcting for multiple testing). Women spoke sentences of social content with lower mean F0, lower F0 variation, lower maximum F0, and with higher intensity variation compared to sentences of neutral content.

In this study, we took great care in scheduling cycle-dependent recording sessions and in standardising the recording procedure. The fertile window was determined using LH tests and confirmed with hormone assays of saliva. We used meaningful sentences, half of which implied a social context and half were of neutral content and asked listeners to indicate how much they would like to get to know the speaker. By using rigorous methodology we found that raters were more interested in getting to know the speakers when their voices were recorded during the fertile late follicular phase compared to during the luteal phase, but only when the speaker spoke sentences implying social interactions (e.g., dating context) and not in neutral sentences. This suggests that menstrual cycle shifts in women's voices might be driven by a higher motivation to meet other people (and in particular men) during days of high fertility.

Data availability: The data associated with this research are available as supplementary material.

Declarations of interest: none

Acknowledgements: The authors wish to thank Amina Bachmann, Sabrina Beeler, Vera Bergamaschi, Gina Camenzind, Emilia da Costa, Fion Emmenegger, Rahel Gfeller, Nik Hunziker, Claudia Ramseier, and Arjeta Velii for their help in coordinating the voice recordings, preparing the voice samples and testing participants.

Funding: This work was supported by the Swiss National Science Foundation awarded to JSL [grant number PP00P1_139072]. The funding source played no role in study design, the collection, analysis and interpretation of data, in writing of the report, and in the decision to submit the article for publication.

References

Abend, P., Pflüger, L.S., Koppensteiner, M., Coquerelle, M., Grammer, K., 2015. The sound of female shape: a redundant signal of vocal and facial attractiveness. Evol & Hum Behav 36, 174-181.

Amir, O., Kishon-Rabin, L., Muchnik, C., 2002. The effect of oral contraceptives on voice: Preliminary observations. J. Voice 16, 267-273.

AudacityTeam, 2015. Audacity [computer program], 2.1.1 ed.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.

Baird, D.D., Weinberg, C.R., Wilcox, A.J., McConnaughey, D.R., Musey, P.I., 1991. Using the ratio of urinary oestrogen and progesterone metabolites to estimate day of ovulation. Stat. Med. 10, 255-266.

Banai, I.P., 2017. Voice in different phases of menstrual cycle among naturally cycling women and users of hormonal contraceptives. PLoS one 12, e0183462.

Barnes, L., Latman, N., 2011. Acoustic measure of hormone affect on female voice during menstruation. International Journal of Humanities and Social Science 1, 5-10.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1-48.

Boersma, P., Weenink, D., 2018. Praat: doing phonetics by computer [computer program], 6.0.40 ed.

Bryant, G.A., Haselton, M.G., 2009. Vocal cues of ovulation in human females. Biol. Lett. 5, 12-15.

Bullivant, S.B., Sellergren, S.A., Stern, K., Spencer, N.A., Jacob, S., Mennella, J.A., McClintock, M.K., 2004. Women's sexual experience during the menstrual cycle: Identification of the sexual phase by noninvasive measurement of luteinizing hormone. J. Sex Res. 41, 82-93.

Collins, S.A., Missing, C., 2003. Vocal and visual attractiveness are related in women. Animal Behav 65, 997-1004.

Eckert, H., Laver, J., 1994. Menschen und ihre Stimmen [Humans and their voices]. Beltz Psychologie Verlags Union, Weinheim, Germany.

Feinberg, D.R., Jones, B.C., DeBruine, L.M., Moore, F.R., Law Smith, M.J., Cornwell, R.E., Tiddeman, B.P., Boothroyd, L.G., Perrett, D.I., 2005. The voice and face of woman: One ornament that signals quality? Evolution & Hum Behav 26, 398-408.

Feinberg, D.R., Jones, B.C., Law Smith, M.J., Moore, F.R., DeBruine, L.M., Cornwell, R.E., Hillier, S.G., Perrett, D.I., 2006. Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. Horm. Behav. 49, 215-222.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 26, 105-109.

Fink, B., Hugill, N., Lange, B.P., 2012. Women's body movements are a potential cue to ovulation. Pers. Indiv. Differ. 53, 759-763.

Fischer, J., Semple, S., Fickenscher, G., Juergens, R., Kruse, E., Heistermann, M., Amir, O., 2011. Do women's voices provide cues of the likelihood of ovulation? The importance of sampling regime. Plos One 6, e24490.

Fox, J., 2016. Bootstrapping regression models. In J. Fox, Applied regression analysis and generalized linear models (3rd ed, pp. 647-668). Sage, Thousand Oaks, CA.

Fraccaro, P.J., Jones, B.C., Vukovic, J., Smith, F.G., Watkins, C.D., Feinberg, D.R., Little, A.C., DeBruine, L.M., 2011. Experimental evidence that women speak in a higher voice pitch to men they find attractive. J Evolutionary Psychology 9, 57-67.

Hardin, J.W., Hilbe, J.M., 2002. Generalized estimating equations. Chapman and Hall/CRC.

Haselton, M.G., Gangestad, S.W., 2006. Conditional expression of women's desires and men's mate guarding across the ovulatory cycle. Horm. Behav. 49, 509-518.
Haselton, M.G., Gildersleeve, K., 2016. Human ovulation cues. Current Opinion in Psychology 7, 120-125.
Haselton, M.G., Mortezaie, M., Pillsworth, E.G., Bleske-Rechek, A., Frederick, D.A., 2007. Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress. Horm. Behav. 51, 40-45.
Halekoh, U., Højsgaard, S. 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. Journal of Statistical Software 59(9), 1-32. Hughes, S.M., Dispenza, F., Gallup, G.G., 2004. Ratings of voice attractiveness predict sexual behavior and body configuration. Evolution & Hum Behav 25, 295-304.
Hughes, S.M., Farley, S.D., Rhodes, B.C., 2010. Vocal and physiological changes in response to the physical attractiveness of conversational partners. J Nonverbal Behav 34, 155-167.
Hughes, S.M., Harrison, M.A., Gallup, G.G., 2002. The sound of symmetry: voice as a marker of developmental instability. Evolution & Hum Behav 23, 173-180.
Jones, B.C., Feinberg, D.R., DeBruine, L.M., Little, A.C., Vukovic, J., 2008. Integrating cues of social interest and voice pitch in men's preferences for women's voices. Biol. Lett. 4, 192-194.
Jones, B.C., Feinberg, D.R., DeBruine, L.M., Little, A.C., Vukovic, J., 2010. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. Animal Behav 79, 57-62.
Jones, B.C., Hahn, A.C., DeBruine, L.M., 2019. Ovulation, sex hormones, and women's mating psychology. Trends in cognitive sciences 23, 51-62.
Karthikeyan, S., Locke, J.L., 2015. Men's evaluation of women's speech in a simulated dating context: Effects of female fertility on vocal pitch and attractiveness. Evolutionary Behav Sci 9, 55-67.
Kowalski, J., Tu, X.M., 2007. Modern Applied U-Statistics. John Wiley & Sons, Hoboken, NJ.
Kuznetsova A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software 82(13), 1-26.
Lobmaier, J.S., Bachofner, L.M., 2018. Timing is crucial: Some critical thoughts on using LH tests to determine women's current fertility. Horm Behav 106, A2-A3.
Meurer, E.M., Garcez, V., von Eye Corleta, H., Capp, E., 2009. Menstrual cycle influences on voice and speech in adolescent females. J. Voice 23, 109-113.
Molenberghs, G., Verbeke, G., 2005. Models for discrete longitudinal data. Springer, New York.
Morris, J. S. (2002). The BLUPs are not 'best' when it comes to bootstrapping. Statistics & Probability Letters, 56, 425-430. doi:10.1016/S0167-7152(02)00041-X.
Nash, J. C., Varadhan R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. Journal of Statistical Software, 43(9), 1-14.
Peirce, J.W., 2007. PsychoPy - Psychophysics software in Python. J. Neurosci Methods 162, 8-13.
Pipitone, R.N., Gallup, G.G., Jr., 2008. Women's voice attractiveness varies across the menstrual cycle. Evolution & Hum Behav 29, 268-274.
Puts, D.A., Bailey, D.H., Cárdenas, R.A., Burriss, R.P., Welling, L.L.M., Wheatley, J.R., Dawood, K., 2013. Women's attractiveness changes with estradiol and progesterone across the ovulatory cycle. Horm & Behav 63, 13-19.
Rehman, R., Khan, R., Baig, M., Hussain, M., Fatima, S.S., 2014. Estradiol progesterone ratio on ovulation induction day: a determinant of successful pregnancy outcome after intra cytoplasmic sperm injection. Iranian Journal of Reproductive Medicine 12, 633-640.

Röder, S., Fink, B., Jones, B.C., 2013. Facial, olfactory, and vocal cues to female reproductive value. Evolutionary Psychology 11, 392-404.
Roney, J.R., 2018. Hormonal mechanisms and the optimal use of luteinizing hormone tests in human menstrual cycle research. Horm Behav 106, A7-A9
Roney, J.R., Simmons, Z.L., 2013. Hormonal predictors of sexual motivation in natural menstrual cycles. Horm. Behav. 63, 636-645.
Roney, J.R., Simmons, Z.L., 2016. Within-cycle fluctuations in progesterone negatively predict changes in both in-pair and extra-pair desire among partnered women. Horm Behav 81, 45-52.
Schneider, B., Cohen, E., Stani, J., Kolbus, A., Rudas, M., ... 2007. Towards the expression of sex hormone receptors in the human vocal fold. J. Voice 21, 502-507.
Schweinberger, S.R., Kawahara, H., Simpson, A.P., Skuk, V.G., Zäske, R., 2014. Speaker perception. WIREs Cogn Sci 5, 15-25.
Shirazi, T.N., Puts, D.A., Escasa-Dorne, M.J., 2018. Filipino women's preferences for male voice pitch: Intra-individual, life history, and hormonal predictors. Adaptive Human Behavior and Physiology 4, 188-206.
Shoup-Knox, M.L., Pipitone, R.N., 2015. Physiological changes in response to hearing female voices recorded at high fertility. Physiol & Behav 139, 386-392.
Snijders, T.A.B, Bosker, R.J., 2012. Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London: Sage.
van Stein, K.R., Strauß, B., Brenk-Franz, K., 2019. Ovulatory Shifts in Sexual Desire But Not Mate Preferences: An LH-Test-Confirmed, Longitudinal Study. Evolutionary Psychology, 17, https://doi.org/10.1177/1474704919848116
Wells, T., Baguley, T., Seargeant, M., Dunn, A., 2013. Perceptions of Human Attractiveness Comprising Face and Voice Cues. Arch Sex Behav 42, 805-811.
Wheatley, J.R., Apicella, C.A., Burriss, R.P., Cárdenas, R.A., Bailey, D.H., Welling, L.L., Puts, D.A., 2014. Women's faces and voices are cues to reproductive potential in industrial and forager societies. Evol & Hum Behav 35, 264-271.

Figure Captions:

Figure 1. Listeners' interest (mean ratings) in getting to know the speakers depending on speakers' menstrual cycle phase and speech content. Error bars represent standard errors.

Figure 2. Random slopes of menstrual cycle phase effects (conditional effects for neutral and social sentences). To obtain the random slopes themselves (rather than the residuals) the respective fixed conditional effect (neutral sentences: 0.05, social sentences: 1.54) was added to each random slope residual.